The PSYCHOLOGICAL RECORD . . . .

NOVEMBER, 1939
Vol. III No. 19

ITEM VALIDITY BY THE ANALYSIS OF VARIANCE: AN OUTLINE OF METHOD

Kenneth H. Baker

THE PRINCIPIA PRESS, INC.
BLOOMINGTON, INDIANA

Price of this number, 15 cents

ITEM VALIDITY BY THE ANALYSIS OF VARIANCE: AN OUTLINE OF METHOD

By Kenneth H. Baker
University of Minnesota

This paper is intended as an aid to the beginner who finds that the sufficient statistic for his problem is an analysis of the variance of his data. Although Snedecor has supplied a very useful handbook[1], most of the problems which he works out are agricultural in detail. The present writer has found that many students of psychology and educational psychology have some trouble in translating these methods which have been worked out with agricultural problems into concepts and procedures which apply to their own research situations. Lev[2] has shown how multiple choice items for use in a scale may be validated and weighted by an analysis of variance, but he does not list the steps in detail. Since Lev's problem is rather more complex than many beginners will feel competent to handle, the writer feels that a simple problem might help more to introduce the student to methods of analysis which are proving to be of greater and greater value in the treatment and interpretation of the results of psychological research.

An analysis of variance is most readily performed when the data are in the form of Table I. Variations of this table may be used for both the single and double criterion of classification and for an equal or unequal number of entries in each of the classes. The entries in the cells of this table may be mean scores, sums of raw scores, proportions or enumeration data. "Classes" refers to one criterion of classification and "Groups" to another. For example, "Classes" might refer to teaching methods and "Groups" to sections of students. For more complete discussions of the type of

[1] Snedecor, G. W., Calculation and Interpretation of Analysis of Variance and Covariance. Ames, Iowa: Collegiate Press, 1934, Pp. 96.

[2] Lev, J., Evaluation of test items by the method of analysis of variance, *Jour. Educ. Psychol.*, 1938, *29*, 623-630.

TABLE I

	Classes				
Groups	1	2	3	4	5
1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
4	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

data to which this method is applicable, the reader is referred to Fisher[3], Snedecor[4], Goulden[5], and others.

In the present problem, "X" is the proportion of individuals passing a certain item in an examination. "Classes" refers to subdivisions of the criterion which was the distribution of the sum of the scores on four similar examinations. Grade in the course might have been used as the criterion, but it was rejected since it was a composite evaluation of several kinds of performance. Total score in examinations represents a criterion more closely related to the validity of items composing the examinations. The distribution of total scores was arbitrarily divided into five criterion classes. Class 1 contained the upper 10% of the distribution; Class 2 the next 15%; Class 3 the next (middle) 50%; Class 4 the next 15%; and Class 5 the lowest 10%. It may be noted that these divisions correspond roughly to the divisions used when letter-grades of "A," "B," etc. are assigned. The proportion of students passing a given item in each of these classes could be calculated, and a valid item would appear as one in which a decreasing proportion of passes occurred in each of the criterion classes. The standard error of the propor-

[3] Fisher, R. A., Statistical Methods for Research Workers. London: Oliver and Boyd, 1936, Pp. 339, Chapter VII. *also* The Design of Experiments. London: Oliver and Boyd, 1937, Pp. 258, Chapters III and IV.

[4] Snedecor, G. W., Statistical Methods. Ames, Iowa: Collegiate Press, 1937, Pp. 341, Chapters X and XI.

[5] Goulden, C. H., Methods of Statistical Analysis. New York: John Wiley and Sons, 1939, Pp. 277, Chapter XII.

tions calculated in this way would be a function entirely of the size of the proportion and the number of cases in each class.

In a problem of this kind, the size of the proportion should have nothing to do with determining the extent of the error of observation. A proportion, for instance, of 90% passing is no more or less significant as such than another proportion of, let us say, 20%. The illusion of the greater reliability of the larger proportion develops from the statistics of inverse probability. These statistics are not strictly applicable to the present problem. A better way to determine the error of the observed proportion is to make several estimates of the proportion and calculate the error of these estimates in terms of their variance.

In the present problem, the five criterion classes are subdivided into four groups each. There are two restrictions to be observed in the subdividing: (1) It must be strictly a randomizing process, and (2) the subdivisions should not contain fewer than ten cases. The first condition can be met easily by using dice or some other equally effective method. When there are fewer than 40 cases in a given criterion class, the second condition can be met by using fewer than four groups for that class. When this is done, a slight correction must be introduced into the procedure to be outlined below. The net result of this method is to give four (instead of one) estimates of the population parameters. The error of observation for the four proportions taken together will, of course, be less than when only one proportion is calculated.

In passing, it should be noted that some problems will require that the scores of those successfully answering an item should be used rather than a simple enumeration of the number passing. The X's in Table I would then refer to a mean score of those passing an item, whereas in the present problem they refer to a proportion of passes. For the present purposes, proportion of passes is adequate. If, however, it is desirable to weight an item (in a scale, for instance), mean score should be used.

To calculate and compare variances, the following steps are recommended. When several items are to be validated, the data of Table I are put on ordinary 3" x 5" or 4" x 6" cards. Table Ia represents such an arrangement, the numbers preceded by the letter "C" representing the various steps in calculation.

TABLE Ia

X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	C 11*	C 12*	
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	C 13	C 14	
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	C 15	C 16	
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	C 17	C 18	
C 1	C 3	C 5	C 7	C 9	C 19		
C 2	C 4	C 6	C 8	C 10		C 20	
C 22	C 23	C 24	C 25	C 26			C 21

$C\,1 = X_{11} + X_{21} + X_{31} + X_{41}$

$C\,2 = X^2_{11} + X^2_{21} + X^2_{31} + X^2_{41}$

$C\,3 = X_{12} + X_{22} + X_{32} + X_{42}$

$C\,4 = X^2_{12} + X^2_{22} + X^2_{32} + X^2_{42}$

$C\,5 = X_{13} + X_{23} + X_{33} + X_{43}$

$C\,6 = X^2_{13} + X^2_{23} + X^2_{33} + X^2_{43}$

$C\,7 = X_{14} + X_{24} + X_{34} + X_{44}$

$C\,8 = X^2_{14} + X^2_{24} + X^2_{34} + X^2_{44}$

$C\,9 = X_{15} + X_{25} + X_{35} + X_{45}$

$C\,10 = X^2_{15} + X^2_{25} + X^2_{35} + X^2_{45}$

$C\,11 = X_{11} + X_{12} + X_{13} + X_{14} + X_{15}$

$C\,12 = X^2_{11} + X^2_{12} + X^2_{13} + X^2_{14} + X^2_{15}$

$C\,13 = X_{21} + X_{22} + X_{23} + X_{24} + X_{25}$

$C\,14 = X^2_{21} + X^2_{22} + X^2_{23} + X^2_{24} + X^2_{25}$

$C\,15 = X_{31} + X_{32} + X_{33} + X_{34} + X_{35}$

$C\,16 = X^2_{31} + X^2_{32} + X^2_{33} + X^2_{34} + X^2_{35}$

$C\,17 = X_{41} + X_{42} + X_{43} + X_{44} + X_{45}$

$C\,18 = X^2_{41} + X^2_{42} + X^2_{43} + X^2_{44} + X^2_{45}$

$C\,19 = C\,1 + C\,3 + C\,5 + C\,7 + C\,9$

$\quad = C\,11 + C\,13 + C\,15 + C\,17$ (check)

$C\,20 = C\,19^2$

$C\,21 = C\,20 \div 20$

$C\,22 = C\,1 \div 4$

$C\,23 = C\,3 \div 4$

$C\,24 = C\,5 \div 4$

$C\,25 = C\,7 \div 4$

$C\,26 = C\,9 \div 4$

* On this page as on following pages the letter C, followed by a figure, represents a step in calculation.

The variances may be summarized in the following table on the reverse side of each card.

TABLE II

Source of variation	Sum of squares	Degrees of freedom	Mean square (variance)	"F"
Total	C 27	19		
Between classes	C 28	4	C 32	C 36
Between groups	C 29	3	C 33	
Error	C 30	12	C 34	C 37
Error	C 31	15	C 35	C 38
Fiducial limits C 39				

$$C\,27 = C\,2 + C\,4 + C\,6 + C\,8 + C\,10 - C\,21$$
$$= C\,12 + C\,14 + C\,16 + C\,18 - C\,21 \text{ (check)}$$
$$C\,28 = \frac{C\,1^2 + C\,3^2 + C\,5^2 + C\,7^2 + C\,9^2}{4} - C\,21$$
$$C\,29 = \frac{C\,11^2 + C\,13^2 + C\,15^2 + C\,17^2}{5} - C\,21$$
$$C\,30 = C\,27 - C\,28 - C\,29$$
$$C\,31 = C\,30 + C\,29$$
$$C\,32 = C\,28 \div 4$$
$$C\,33 = C\,29 \div 3$$
$$C\,34 = C\,30 \div 12$$
$$C\,35 = C\,31 \div 15$$
$$C\,36 = C\,32 \div C\,33 \quad \text{(or vice versa, if C 33 is larger)}$$
$$C\,37 = C\,34 \div C\,33 \quad \text{(or vice versa, if C 33 is larger)}$$
$$C\,38 = C\,32 \div C\,35 \quad \text{(or vice versa, if C 35 is larger)}$$
$$C\,39 = \sqrt{\frac{2 \times C\,35}{4}} \times 3.71$$

Notes on calculations: The operations outlined here are arranged for machine calculation. Calculations C 1 through C 18 may be done in pairs, i.e., C 1 with C 2, C 3 with C 4, etc.

The check for calculation C 19 tells only of the correctness of

the total for the rows and columns. The check for the sums of squares is not made until calculation C 27, although this is seldom in error, if the calculator is operating correctly.

Calculation C 21 gives the correction factor. The "20" in this operation refers, of course, to the number of cells in the original table. It does not refer to the number of cases or operation C 20.

Calculations C 22 through C 26 are for the means of the columns. In a complete analysis of a double criterion of classification, means of the rows would also be calculated. The latter means are obviously of no value in the present problem.

Before the division is performed in calculations C 28 and C 29, the item-count dial of the calculator should show a figure equal to C 19.

Calculations C 30 and C 31 are obtained by subtraction and will be incorrect if previous calculations are inaccurate. It would seem that the checks introduced so far are adequate to minimize the possibility of error.

Calculation C 36 gives an "F" which tells whether the variance between the criterion classes is significantly greater than the variance within these classes.

Calculation C 37 gives an "F" which tells whether the variance attributable to the randomization process is significantly different from the variance from other sources. This is a rough check on the adequacy of the size of the sample upon which each of the proportions is based. In a strict analysis of a double criterion of classification, this "F" tells whether the group means come from a homogeneous population of means.

Calculation C 38 gives the most important "F" for an item validity study since the "Error" here includes the variance attributable to all sources other than the criterion classification.

Calculation C 39 gives the amount by which the column means must vary in order to satisfy the 1% level of significance for the appropriate (in this case, six) degrees of freedom. The calculation is based on the assumption that the variances of each of the classes do not differ significantly, i.e., that the variance of the variances is low. A rough check of the truth of this assumption may be made by applying a chi-square test: Chi-square $= \frac{\Sigma V^2 - (\Sigma V)^2/n}{m}$

where m is the theoretical variance of the total population the best estimate of which is $\frac{2\ a^4}{k\text{-}1}$ in which a^4 is the square of the mean of the variances within the classes, k the number in the groups. When the number in the groups is unequal, k can be estimated by taking the harmonic mean. This test of the homogeneity of the variances is usually not necessary, although it should be made, if a thorough analysis is to be carried out.

In conclusion, it must be pointed out that, although this method of analysis milks the data of every bit of information they contain, it is impractical from the standpoint of time unless the items to be validated are to be used in some sort of scale. To use this method in the validation of items in an ordinary class room examination requires more time than the results of the analysis would justify. On the other hand, for the study of items in questionnaires, scales and other research instruments, the analysis of variance is most effective.